Constructing a broad-coverage lexicon for text mining in the patent domain

Nelleke Oostdijk[*], Suzan Verberne[*], Cornelis Koster[†]
[*]Information Foraging Lab & Centre for Language and Speech Technology
[†]Information Foraging Lab and Dept. of Computer Science
Radboud University Nijmegen
P.O. Box 6500 HD Nijmegen, The Netherlands
E-mail: {n.oostdijk|s.verberne}@let.ru.nl, kees@cs.ru.nl

Abstract

For mining intellectual property texts (patents), a broad-coverage lexicon that covers general English words together with terminology from the patent domain is indispensable. The patent domain is very diffuse as it comprises a variety of technical domains (e.g. Human Necessities, Chemistry & Metallurgy and Physics in the International Patent Classification). As a result, collecting a lexicon that covers the language used in patent texts is not a straightforward task. In this paper we describe the approach that we have developed for the semi-automatic construction of a broad-coverage lexicon for classification and information retrieval in the patent domain and which combines information from multiple sources. Our contribution is twofold. First, we provide insight into the difficulties of developing lexical resources for information retrieval and text mining in the patent domain, a research and development field that is expanding quickly. Second, we create a broad coverage lexicon annotated with rich lexical information and containing both general English word forms and domain terminology for various technical domains.

1. Introduction

That lexical information is indispensable for realistic natural language processing (NLP) systems is well-known. Already in 1996 Boguraev and Pustejovsky observed that "regardless of a system's sophistication or breadth, its performance must be measured in large part by the resources provided by the computational lexicon associated with it" (p. 3).

The challenge in the 1980s and 1990s consisted in the scaling up to size of NLP prototype applications and it was in this context particularly that the issue of lexical coverage was raised. There was an urgent need for extending the lexical coverage, but it was equally important to improve on the richness of the linguistic information.[1] During these years much effort was directed at investigating what (and how) information could be successfully extracted from machine-readable dictionaries (MRDs) and text corpora. Ide and Véronis (1994), taking stock of the state of the art in the mid-1990s, summarize the situation as follows:

> "it is now widely recognized that knowledge base construction requires combining information from multiple resources, especially information provided by corpus analysis, since corpora can provide information such as common collocates, proper nouns, role preference information, frequency of use and similar statistics, etc. However, with corpora as with MRDs, fully automatic extraction is not likely, and it is again unclear what corpora can provide and how valuable the information is for NLP."

In later years the issue of lexical knowledge acquisition for use with NLP applications remained unsettled.

In this paper we describe the approach that we have developed for the semi-automatic construction of a broad-coverage lexicon for text mining in the patent domain and which combines information from multiple sources.

The structure of the paper is as follows: After a description of the text mining system and the lexical information that is needed, we describe the set-up that we have chosen for acquiring and maintaining the necessary lexical information. Section 3 introduces the lexical database and describes the construction of the initial base lexicon. Section 4 elaborates on the pipelined procedure that we have developed for expanding the lexicon with domain terminology. In Section 5 we evaluate what have achieved in terms of lexical coverage and reflect on the effectiveness of our procedure. Section 6 summarizes the work described in the paper and also the work that is foreseen for the future. The paper is a report on work in progress and aims to share the experience and insights gained so far.

2.Text Mining for Intellectual Property

2.1 Text mining system

In the Text Mining for Intellectual Property (TM4IP) project we are implementing a text mining system for intellectual property search (Koster et al., 2009). The system consists of (1) an English hybrid dependency parser (AEGIR) that is especially developed for use in the patent domain, and (2) a professional search engine that uses structured queries based on dependency relations (Koster et al., 2006).

AEGIR combines syntactic rules with an extensive word form lexicon (the parser lexicon) and information about the frequencies of deep syntactic (dependency) relations between words. This information is stored in a database of dependency triplets (the triplet database) and is consulted during the parsing process. Figure 1 gives a schematic representation of the different components that make up the parser.[2]

[1] Cf. Boguraev (1991: 164)

[2] AGFL is a parser generator that was developed at the

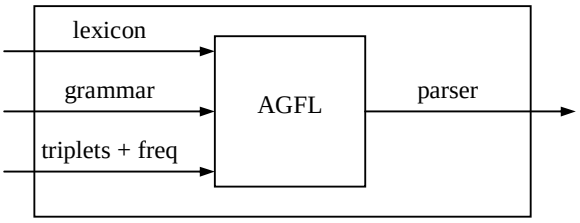

Figure 1: Parser generation

For application in a text mining system for the patent domain, the strength of AEGIR is its capacity to effectively perform normalization at various levels, viz. the levels of typography (e.g. upper and lower case, spacing), spelling (e.g. British and American English, hyphenation), morphology (lemmatization of word forms) and syntax (standardization of the word order, for example by transforming passive structures into active ones).

2.2 Lexicon and triplet database

Together the lexicon and the triplet database (henceforth referred to as TDB) should contain all lexical information that is required by the parser. In the lexicon all word-related information is stored, together with lexical frequencies for those words that can have more than one part of speech. The TDB holds information pertaining to the frequency of dependency relations between lemmas.

In the lexicon we only include non-compositional lexical entries. Thus the size of the lexicon is minimized. We expect the parser to handle all compositional multi-word entries (e.g. *internal combustion*, *carbon atoms*, *light-emitting*) as well as any compositional complex single-word entries (such as chemical formulae). Certain items — such as numerals, names, and dates — are not included in the lexicon. Instead they are dealt with by a grammar component which allows for their robust recognition. Words that the parser encounters that are not in the lexicon and which cannot be recognized robustly, are ultimately skipped.

The lexicon consists of two types of file: the .dat files which include all word forms together with their formal properties and the .fct files which include all lemmas together with the relevant subcategorization information. For example, the entry for the word form 'consists' in the .dat file is as follows: [3]

"consists" V("consist",sing,third) 2548

while in the .fct file we find

"consist" verbsel("consist",none,intr,,of)
"consist" verbsel("consist",none,intr,,with)
"consist" verbsel("consist",none,intr,in)

The TDB supplements the lexicon information as it includes previously observed, reliable dependency triplets. Examples of triplets in the TDB involving "consist" are[4]

12 [examination,SUBJ,consist]
14 [series,SUBJ,consist]
10 [consist,PREPof,element]
5 [consist,PREPof,layer]
9 [consist,PREPof,part]

The TDB is dynamic in the sense that it expands and develops over time. Reliable dependency triplets can be harvested from lexical resources such as thesauri (see below) and from treebanks. Further dependency triplets may be obtained as a result of a bootstrapping process, using the AEGIR parser itself for parsing texts.

2.2.1 Lexical normalization

In order to maximize the findability of terms in search and retrieval we use lemmas rather than word forms. For the same reason we normalize various spelling variants. Thus in the lexicon word forms involving spelling variants are associated with a single, default lemma. For example,

"rigour" N("rigour",sing)
"rigor" N("rigour",sing)

2.2.2 Lexical coverage

The lexical information that is needed for mining intellectual property texts (patents) comprises both general English vocabulary and vast quantities of domain terminology. The patent domain is very diffuse as it comprises a great many technical fields, ranging from Human Necessities to Chemistry & Metallurgy and from Engineering to Biomedicine. As a result, collecting a lexicon that covers the language used in patent texts is not a straightforward task. In Sections 3 and 4 we describe our approach.

3. Building the Lexicon: Set-up

3.1 Approach

For collecting lexical data for the kind of broad-coverage lexicon described in the previous section, we use a variety of resources: computer-readable lexicons, thesauri, treebanks and text collections. With the help of automatic processes we extract as much information from these resources as we can. The role of the human expert is restricted to the following tasks: (1) specifying how different annotation/classification schemes should be mapped onto each other; (2) resolving ambiguous cases; (3) providing information where this is found to be missing; (4) correcting errors.

In building the lexicon we distinguish between the parser lexicon on the one hand and the lexical database underlying the lexicon on the other hand. In the lexical database, lemmas are stored with all relevant lexical information such as subcategorization. The parser lexicon itself is a word form lexicon in which the lemma and lexical frequency of each word are stored with the word form. Each new version of the parser lexicon is

Radboud University Nijmegen. For a description see Koster et al. (2006) and also http://www.agfl.cs.ru.nl/

[3] Since the .dat and .fct files are AGFL files (just like the files containing the grammar rules) they conform to the notional conventions of the AGFL formalism. Thus V and verbsel are considered to be non-terminals.

[4] The notation reads as follows: between *examination* and *consist* a SUBJ(ective) relation holds such that *examination* is the subject of *consist*. The numbers denote frequencies.

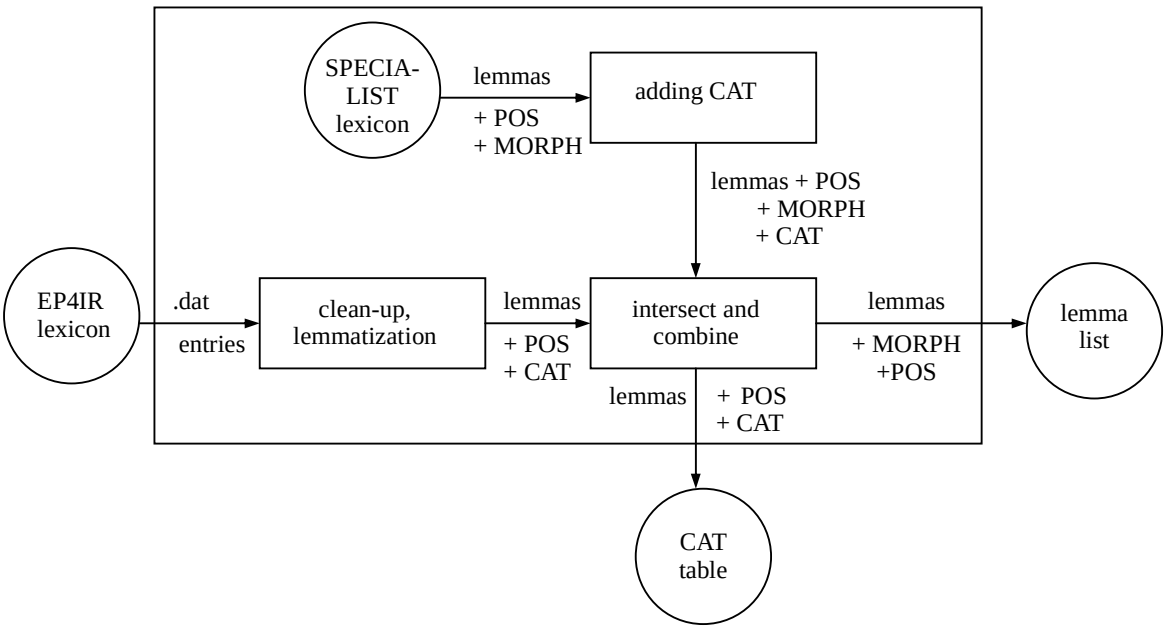

Figure 2: Construction of the initial base lexicon

generated from the lexical database on the basis of a set of rules. The reasons for using a parser lexicon containing word forms and a separate underlying database containing lemmas are that (1) lexical look-up by the parser is much faster than morphological analysis and (2) a lexical database of lemmas makes maintenance of the lexicon easier.

3.2 The lexical database

The lexical database follows the example set by the SPECIALIST lemma lexicon[5] (Browne, 2000), storing lemmas rather than word forms and including information as to how lemmas can be expanded into word forms in order to generate the parser lexicon. In our database we distinguish between lexical items belonging to the open classes (viz. noun, adjective, adverb and verb) and items from the closed classes (such as conjunctions, pronouns and determiners). The reason for making this distinction is that the set of closed class items is a stable set: it suffices to decide once which items are involved and what information is associated with them; after that the set is no longer subject to change.[6] In what follows we shall focus on the open classes.

The information associated with the four open-class lemma types in our database is as follows:

- NOUN: lemma, part of speech (POS), inflection, countability, noun type, subcategorization
- ADJECTIVE: lemma, POS, inflection, adjective type, subcategorization
- ADVERB: lemma, POS, inflection, adverb type
- VERB: lemma, POS, inflection, verb type, subcategorization, verb particle

From the point of view of building a lexical database that can be re-used in a wider range of applications (rather than just the current project) and also for ease of maintenance, it makes sense to distinguish between the different types of information. Thus, there is the stable formal information, viz. lemma, POS and inflection. In addition, there is information that may vary, depending on the linguistic descriptive framework adopted (type, subcategorization), and that will develop over time as observed usage brings to light previously unattested occurrences. While the former type of information is stored as part of the lemma list, the latter is included in the CAT Table. In generating the lexicon the lemma list is used to produce the .dat files, while information from the CAT Table is incorporated in the .fct files (cf. Figure 2).

3.3 The base lexicon

As a first step towards populating our lexical database we used the lexicon that was constructed in the EP4IR project (Koster and Verbruggen, 2002) From this lexicon we extracted all single words (approximately 575,000 word forms).7 For the purpose of checking their validity, we compared these to the entries in the SPECIALIST

[5] The SPECIALIST Lexicon has been developed to provide the lexical information needed for the SPECIALIST Natural Language Processing (NLP) System. It is intended to be a general English lexicon that includes many biomedical terms. Coverage includes both commonly occurring English words and biomedical vocabulary. The lexicon entry for each word or term records the syntactic, morphological, and orthographic information needed by the SPECIALIST NLP System. http://lexsrv3.nlm.nih.gov/SPECIALIST

[6] Closed class items are also different from the open class items in the sense that their classification and whatever additional information is provided heavily depends on the descriptive framework adopted and therefore the requirements made by the grammar underlying the parser.

[7] A *single word* is defined here as a word without a hyphen or a blank space.

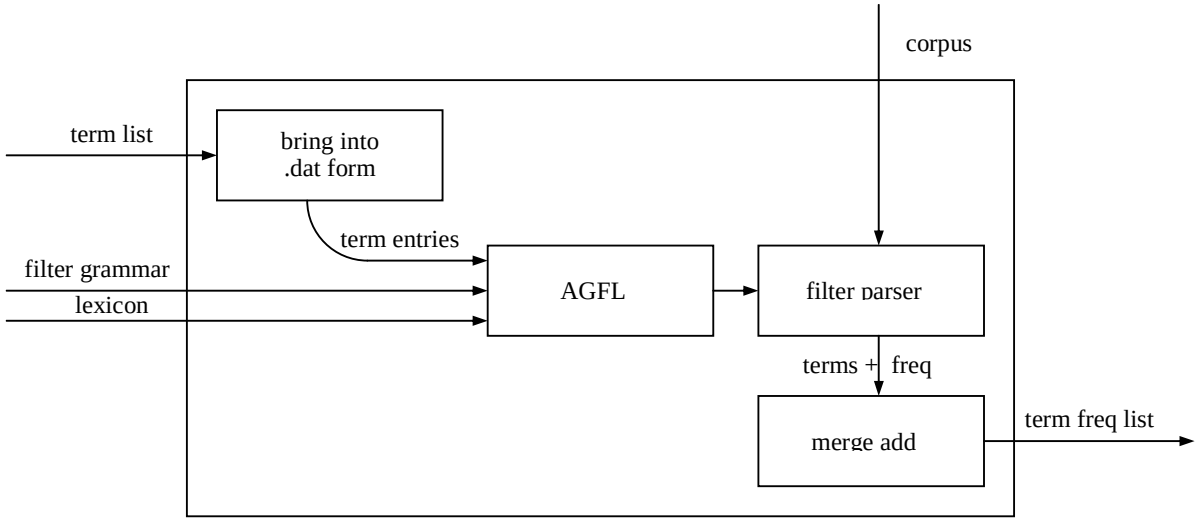

Figure 3: Filtering

lexicon. Thus we extracted from the SPECIALIST all single-word entries and the information associated with these items. As a dedicated lexicon, the SPECIALIST includes approximately 12,000 general English lemmas and 20,000 lemmas from the English biomedical domain. The information was then mapped onto the targeted format. We wrote a set of morphological expansion rules to generate all possible word forms. As a result we obtained approximately 250,000 unique word forms in the open classes. The intersection of the expanded SPECIALIST and the EP4IR lexicon provided us with an initial high quality base lexicon of approximately 220,000 word forms. The lemmas associated with the word forms in the intersection were stored in the database. Entries that were not part of the intersection were passed onto the human expert for manual inspection.

4.Large-scale Acquisition

As mentioned above (Section 2.2.2), patent documents contain domain terminology from several technical and (bio)medical fields. In order to increase the coverage of our parser for domain terminology, we expand our initial English base lexicon by harvesting external resources such as glossaries, terminologies, thesauri and text collections. In this section we describe the procedure that we have developed for the acquisition of large quantities of lexical data (domain terminology) with a minimum of human intervention. Before we describe the mono-word and multi-word pipelines in Section 4.2, in the next section we first introduce one of the main tools: the filter program.

4.1 Filtering

Many tasks directed at the extraction of lexical data from existing resources involve some kind of filtering. Filtering is done by means of filter programs. These are dedicated parsers which are based on a rule-based filter grammar, a lexicon and an incoming term list (cf. Figure 3). In each case the grammar consists of a highly restricted set of rules that focus on the recognition of a particular construct (e.g. NP).

Application of a filter program to a corpus or some other text collection will yield a list of candidate lexical entries with their corpus frequency.

4.2 The mono-word and multi-word pipelines

The procedure that we have developed for the large-scale acquisition of lexical entries is depicted in Figures 4 and 5. It involves two pipelines: the mono-word pipeline and the multi-word pipeline. In the next two sections we first describe these pipelines. In section 4.2.3 we illustrate how they were used to harvest lexical entries from the UMLS.

4.2.1 Multi-word pipeline

Given a terminology list (e.g. a thesaurus, glossary or domain lexicon) we first identify the words that do not yet appear in the current version of our AEGIR lexicon (cf. Figure 4). After removing all known words, we also remove any suspicious items, such as terms containing punctuation and other non-word characters The remaining items then are input for a filter program which yields a frequency list of candidate lexical entries (cf. Section 4.1). Note that we disregard terms from the terminology list that do not occur in a large domain corpus. Single-word items are passed onto the mono-word pipeline straightaway; multi-word items undergo compositional analysis. Where items are found to be compositional we obtain a set of triples that we store in the TDB. Non-compositional items are passed onto the human expert who has to approve them. Those items that are approved are passed onto the mono-word pipeline for further processing.[8]

[8] The term *mono-word* is used here to denote single words (cf. Note 7) as well as non-compositional multi-words.

4.2.2 Mono-word pipeline

The mono-word pipeline is a procedure that is used to provide the necessary information with newly acquired words.[9] The human expert works his way through the list of new words, starting with the most frequent ones, adding the POS and morphological information as appropriate. Next the words are lemmatized. The lemmas, POS and information on how to expand the lemmas into word forms are then stored in the lexical database (lemma list). After adding subcategorization to the lemmas, this information is stored in the CAT table.

4.2.3 Harvesting lexical entries: An example

In this section, by way of illustration, we demonstrate the use of the lexical pipelines. We describe the complete procedure that we followed for harvesting lexical entries from one specific resource: the UMLS Metathesaurus of medical terminology (Bodenreider, 2004). The input of our lexical pipeline is the complete list of 1.26 Million terms from the UMLS thesaurus without its hierarchical structure. The following filtering processes are now applied to this term list:

Step 1. All terms from the UMLS list that are already covered by the AEGIR lexicon (e.g. *aspirin, animal, anger*) are removed from the term list and ignored in the remainder of the pipeline. 1.24 Million UMLS terms remain after this step.

Step 2. We apply a script to the UMLS term list that removes all items that are considered suspicious (for example because of uneven bracketing or unexpected punctuation; e.g. *aa comb.no3/b/mv-ao4/mn/min aa unidentified*). 1.18 Million UMLS terms remain after this step.

Step 3. We apply a corpus filter to determine the frequency of these terms.

Step 4. We split the remaining 1.18 Million UMLS terms in single-word terms and multi-word terms (terms containing at least one white space or hyphen). 928,000 terms in the list are multi-word terms.

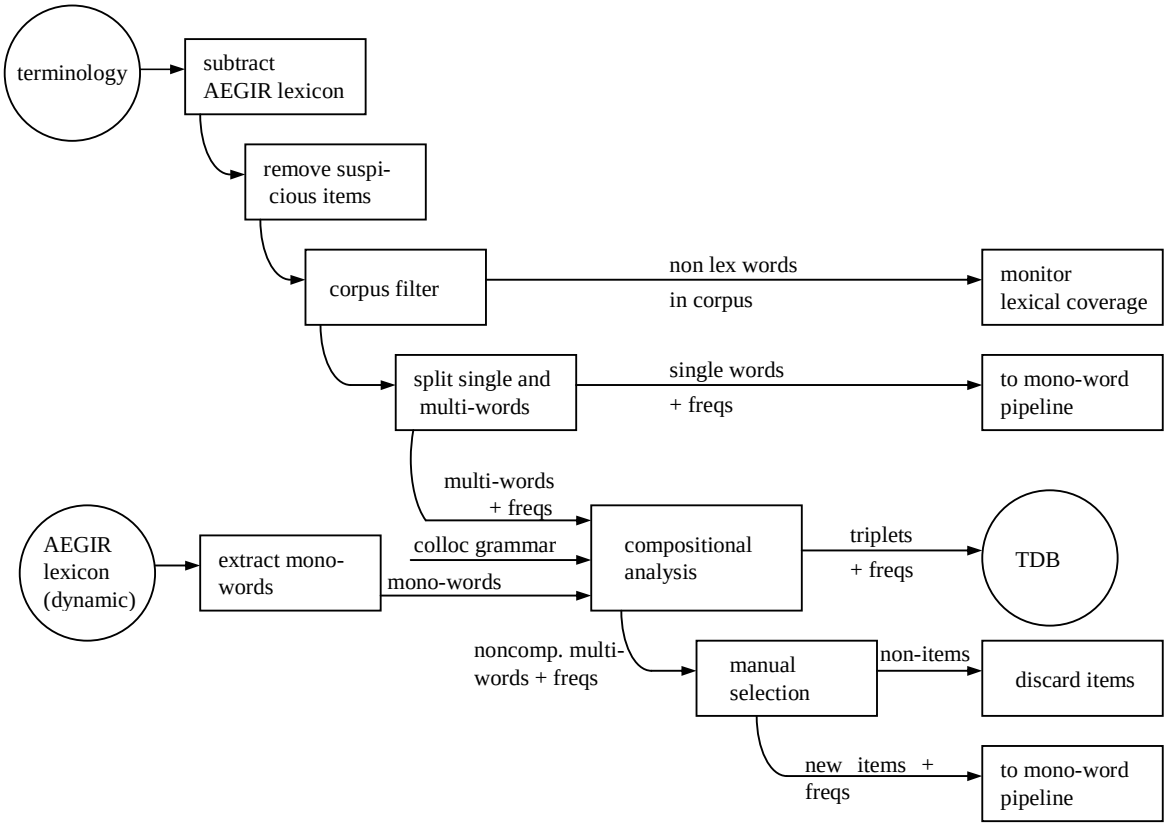

Figure 4: Multi-word pipeline

[9] Although the main purpose of the mono-word lexical pipeline is to process new data from lexical resources, we also used it for cleaning up the initial base lexicon.

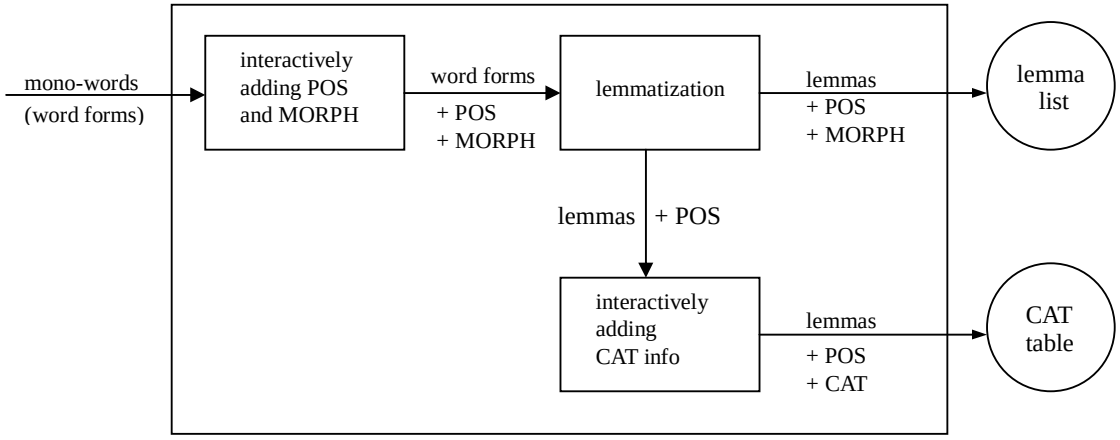

Figure 5: Mono-word pipeline

Step 5. The 928,000 multi-word terms are analyzed by an NP parser in order to determine whether they can be compositionally derived from the mono-word (i.e. single or non-compositional multi-word) entries in the parser lexicon. For example, the UMLS term *Mannich base* cannot be analysed as a compositional phrase because the parser does not know the proper noun *Mannich* as a lexicon term. On the other hand, *amino group* can be analyzed as a compositional phrase with the structure [group,ATTR,amino].

Step 6. All non-compositional multi-word terms are looked up in a sub-corpus of MAREC[10] to determine their corpus frequency. In the UMLS list, only 661 of the non-compositional multi-word terms occur in the corpus (e.g. *chemokine receptor* with frequency 17 and *Mannich base* with frequency 6). The most frequent candidate terms are manually judged, and annotated with essential information for the lexical database.

Step 7. All compositional multi-word terms from the UMLS list are not included in the lexicon because they are appropriately processed by the parser itself. Instead, we transform the multi-word terms that occur at least once in the MAREC sub-corpus (2,376 UMLS terms) to dependency triplets and add them to the TDB, together with their corpus frequencies. For example,

164 [sequence,ATTR,DNA]
287 [group,ATTR,amino]

5. Evaluation

We evaluated the AEGIR lexicon by measuring the lexical coverage of the AEGIR lexicon that we developed on the subset of the MAREC corpus (7 Million words) and we compared it to the CELEX lexicon (Baayen et al., 1993).

For measuring lexical coverage. we used a corpus filter as described in Section 4.1. The corpus filter allows us to skip over special tokens such as single characters, numerals and formulae. Since these tokens are robustly recognized by the AEGIR parser, they should not be included in the lexicon. We measured lexical coverage both on the token level (counting duplicate words separately) and the type level (counting duplicate words once). A type-level count of course gives a lower lexical coverage because the words that are not covered by the lexicon are generally lower-frequency words. The lexical coverage (both type and token counts) for the AEGIR and CELEX lexicons on the MAREC sub-corpus are in Table 1.

lexicon	MAREC sub-corpus coverage	
	types	tokens
AEGIR	86.5%	99.8%
CELEX	60.4%	98.8%

Table 1: Lexical coverage

Table 1 shows that especially in type counts, the AEGIR lexicon has a higher coverage than the CELEX lexicon.

Of all the word forms in the base lexicon, only a small subset (10,051 items) are ambiguous as regards their parts of speech. The lexical frequencies of these words play an important role in disambiguating them in the actual context in which they occur. For example, in patent texts the word *claim* is found to be much more frequent as a noun (N) than as a verb(V):

"claim"	N("claim",sing)	100069
"claim"	V("claim",infi)	5223
"claim"	V("claim",plur,PERS)	3062
"claim"	V("claim",sing,first|secnd)	4305

A problem that we came across while gathering frequency information was that most text corpora do not contain lemma information. Thus while we could quite easily establish what the frequency was of a particular word form with a single lemma, it proved impossible to automatically determine the frequency of those word form-lemma pairs where the word form can be associated with multiple lemmas. Examples of the latter include the following:

[10] MAREC stands for MAtrixware Research Collection, which is a collection of patent documents. Matrixware supplied 400,000 documents from this collection for use in the AsPIRe'10 workshop. Here we use a subset of 7 million words for our frequency counts.

"axes" N("axe",plur)
"axes" N("axis",plur)

"putting" V("put",infi)
"putting" V("putt",infi)

We therefore decided to extract from our lexicon all (836) items that were ambiguous for their lemmas. A human expert was then asked to give an estimate of their relative distribution.

6. Conclusion

In this paper, we have described an approach to developing a broad coverage lexicon containing both general English word forms and domain terminology for various technical fields. At present, we have constructed a first version of the lexical database and the parser lexicon that is generated from it. We have constructed a word form lexicon containing words from the open word classes together with their POS, frequency and subcategorization information. Moreover, we have developed a pipelined procedure for processing terminology lists that we harvest from glossaries, terminologies and thesauri. We applied the lexical pipeline to a number of thesauri (UMLS and WordNet), gaining a set of candidate lexical entries.

Our future work consists of three main tasks: (1) to further expand and improve our parser lexicon and the underlying lexical database, (2) to build and expand the TDB by extracting reliable triplets from treebanks and through a bootstrapping procedure involving the application of the AEGIR parser on corpora and other text collections, and (3) to evaluate the effect of lexical quality and coverage on the accuracy of the AEGIR parser and the professional search engine in which the parser is included.

7. Acknowledgement

The Text Mining for Intellectual Property (TM4IP) project is funded by Matrixware Information Services GmbH, Austria.

8. References

Baayen, R., Piepenbrock, R. & van Rijn, H. (1993). *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. In *Nucleic Acids Research*: 32 (Database Issue), D267--D270.

Boguraev, B. (1991). Building a Lexicon: An Introduction. *International Journal of Lexicography*, Vol. 4 No. 3. Oxford University Press, pp. 163--166.

Boguraev, B. & Pustejovsky, J. (1996). Issues in Text-based Lexicon Acquisition. In B. Boguraev & J. Pustejovsky. (Eds.), *Corpus Processing for Lexical Acquisition.* Cambridge, Mass.: The MIT Press, pp. 1-17.

Browne, A., McCray, A. & Srinivasan, S. (2000). The Specialist Lexicon. In *Library of Medicine Technical Reports*, pp. 18--21

Ide, N. & Véronis, J. (1994). Machine Readable Dictionaries: What have we learned, where do we go? In *Proceedings of the International Workshop on the Future of Lexical Research.* Beijing, China, 137--46.

Koster, C., Oostdijk, N., Verberne, S. & D'hondt, E. (2009). Challenges in Professional Search with PHASAR. In *Proceedings of DIR 2009*, pp. 101--102.

Koster, C., Seutter, M. & Seibert, O. (2006). The Phasar Search Engine. In *Proceedings NLDB 2006.* Springer LNCS 3999. pp. 141--152.

Koster, C., Seutter, M. & Seibert, O. (2007). Parsing the Medline Corpus. In *Proceedings RANLP 2007*, pp. 325--329.

Koster, C. & Verbruggen, E. (2002). The AGFL Grammar Work Lab. In *Proceedings FREENIX/ Usenix*. pp. 13--18.

Verberne, S., D'hondt, E., Oostdijk, N. & Koster, C. (2010). Quantifying the Challenges in Patent Claims Processing. In *Proceedings of the AsPIRe'10 workshop*.